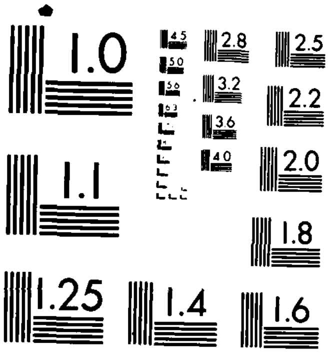
AD-A161 631 AN ELEMENTARY STATISTICAL APPROACH TO MEASURING UNCERTAINTY IN A COST ESTIMATE RANGE(U) MITRE CORP BEDFORD MA P R GARVEY 1985 F19628-84-C-0001 1/1

UNCLASSIFIED F/G 5/1 NL

END

FILMED

DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

An Elementary Statistical Approach To Measuring Uncertainty In A Cost Estimate Range

by

Paul R. Garvey
Technical Staff
The MITRE Corporation
Bedford, MA

Presented at the
19th Annual Department of Defense
Cost Analysis Symposium
Xerox Training Center
Leesburg, Virginia
September 17 - 20 1985

AN ELEMENTARY STATISTICAL APPROACH TO MEASURING UNCERTAINTY IN A COST ESTIMATE RANGE

P. R. Garvey*

The intent of this article is to suggest an approach to developing a probabilistic cost range for a system in the conceptual design phase. This methodology has been applied to a recent software cost study for a large scale military acquisition program, hence the emphasis in this paper will be on the problem of determining the most probable software development cost interval.

INTRODUCTION

In this article we will consider a system as a regularly interacting or interdependent group of items comprised of hardware and/or software elements forming a unified whole.

In many large scale command and control projects, Prime Mission Product (PMP) cost estimates developed for Full Scale Engineering Development (FSED) are usually reported as a range, and hence are not necessarily intended for budgetary purposes, but rather to provide information to the respective Program Office to aid in system engineering trade-off studies and acquisition planning activities.

When a project is in concept definition, or initial development, the precise determination of system cost is usually not possible. For software intensive systems, estimates of Computer Program Configuration Item (CPCI) size prior to FSED are subjective, and are often based on comparable software tasks, or from advanced prototype designs. The variability in a software cost estimate is directly related to the variability in CPCI size, which may vacillate around data points from low = a′, most likely = m′, to high = b′ estimates of lines of code (LOC):

$$\text{LOC}_{\text{range}}: \quad a' < m' < b'.$$

A hierarchical overview of the procedures for developing a software cost range is shown in Figure 1.

*The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policy, either expressed or implied, of the United States government.

This work was sponsored by the Electronic Systems Division, Air Force Systems Command, Hanscom AFB, Bedford, Massachusetts under Contract No. F19628-84-C-0001.

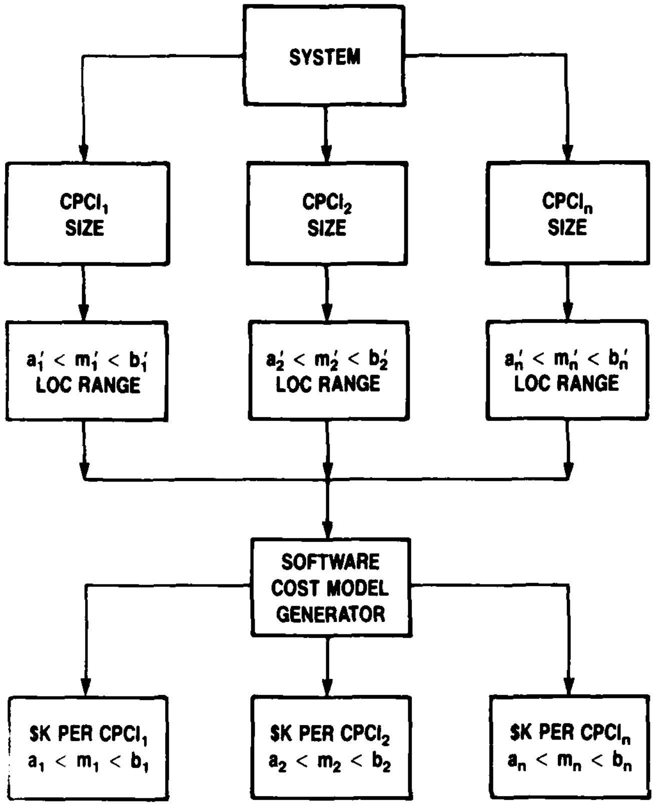

SYSTEM
CPCI$_1$ SIZE
CPCI$_2$ SIZE
CPCI$_n$ SIZE
$a'_1 < m'_1 < b'_1$ LOC RANGE
$a'_2 < m'_2 < b'_2$ LOC RANGE
$a'_n < m'_n < b'_n$ LOC RANGE
SOFTWARE COST MODEL GENERATOR
$K PER CPCI$_1$ $a_1 < m_1 < b_1$
$K PER CPCI$_2$ $a_2 < m_2 < b_2$
$K PER CPCI$_n$ $a_n < m_n < b_n$

Figure 1. System Overview

It is not my intention to address the inherent estimation error associated with many of the parametric, or non-parametric software cost models. Rather, the following discussion attempts to provide a non-rigorous method to measure the degree of uncertainty in a software cost estimate range generated when only subjective technical assessments on LOC are available. Although this probability technique was developed for treating software costs, it can easily be extended to any other costs, such as hardware, which are stated as a range.

THE EXPECTED COST

In circumstances where it becomes necessary to report a system cost as the most representative point in the cost estimate range, a useful measure of central tendency to determine is the expected cost. The expected cost is the point of the center of gravity in the system cost range. Mathematically, the expected cost is defined by

$$\mu = E(X) = \int_I x f(x) dx \quad I: [a \leq x \leq b]$$

where X denotes the cost random variable, and f is the continuous probability density function (pdf) of X. The integral limits a and b represent respectively the low and high extrema in the cost interval I. By definition, f(x) must satisfy the following properties

$$* \ f(x) \geq 0 \text{ in I,} \qquad I: [a < x < b]$$

$$* \int_I f(x)dx = 1 \qquad \text{for the continuous case}$$

Since the true underlying distribution of X is unknown, a probability distribution of cost may be expressed by choosing an appropriate probability function that most accurately reflects the unique system cost behavior. Define $\hat{f}$ as the pdf that is the analyst's "best" subjective approximation to the true underlying density function f. Thus

$$\hat{f} \simeq f$$

We will further require the approximating probability function to satisfy the boundary conditions $\hat{f}(a) = 0$ and $\hat{f}(b) = 0$. The maximum value of $\hat{f}$ is defined by

$$\hat{f}(m) = \max_{x \in I} \hat{f}(x) \qquad I: [a \leq x \leq b]$$

Several classes of probability functions satisfy the above criteria. This article will consider

- A polynomial density function
- A triangular density function

Expressions for their means and variances will be derived.

A Polynomial Density Function

Consider the situation where an analyst obtains subjective values of a and b, but the most likely value m is not given as a point, but as a percentage from the lower bound of I. Define a 4-degree unimodal polynomial density function by g(x) where

$$g(x) = \sum_{n=0}^{4} C_n x^n, \quad I: \quad [a \le x \le b]$$

which also satisfies the conditions that g(a) = 0, and g(b) = 0. Further, assume there exists a unique maximum point m contained in I such that

$$g(m) = \max_{x \in I} g(x)$$

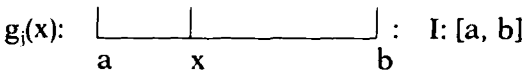
The following discussion considers two distinct functional variations of g(x). These forms, denoted by $g_j(x)$ (j = 1 or 2), are each uniquely determined by the location of their mode m_j (j = 1 or 2), where

$$g_1(m_1) = \max_{x \in I} g(0.3(b-a)+a)$$

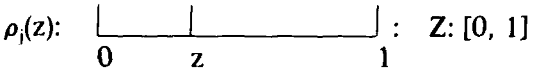
$$g_2(m_2) = \max_{x \in I} g(0.7(b-a)+a)$$

Expressions for the mean and variance of g_j will be derived.

To reduce the computational complexity when computing the mean and variance of g_j, transform the initial interval I



to the unit interval Z



by the linear transformation

$$z = \frac{x-a}{b-a}$$

and form a 4-degree unimodal polynomial density function, $\rho_j(z)$ (j = 1 or 2), on the $Z_{[0,1]}$ interval as shown in Figure 2.

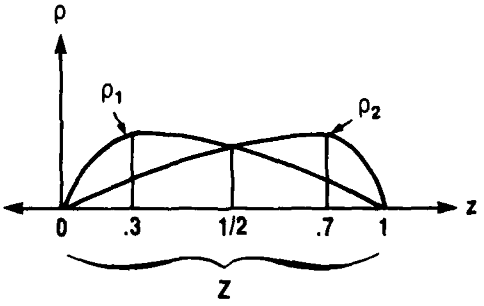


Figure 2. A Polynomial Density Function

The value of $g_j(m_j)$ (j = 1 or 2) when transformed into the unit Z interval occurs at the points

$$z_1 = \frac{(0.3(b-a)+a)-a}{b-a} = 0.3\text{ , or}$$

$$z_2 = \frac{(0.7(b-a)+a)-a}{b-a} = 0.7$$

as shown in Figure 2.

The equations representing ρ_j (j=1 or 2) are

$$\rho_1 = \frac{15}{2}(1-z)^2\left[1-(1-z)^2\right], 0 \le z \le 1$$

$$\rho_2 = \frac{15}{2} z^2(1-z^2), 0 \le z \le 1$$

Note that this density function is symmetric about 1/2, and

$$\rho_1(z) = \rho_2(1-z), 0 \le z \le 1$$

The expected value of ρ_j (j = 1 or 2) is defined by

$$E(Z_j) = \int_Z z\rho_j(z)dz \qquad Z: [0, 1]$$

and the variance σ_Z^2 is computed from

$$\sigma_Z^2 = E(Z_j^2) - E(Z_j)^2$$

where

$$E(Z_j^2) = \int_Z z^2\rho_j(z)dz \qquad Z: [0, 1]$$

On the unit $Z_{[0,1]}$ interval we then have

$$E(Z_1) = \frac{3}{8}, \qquad E(Z_2) = \frac{5}{8}$$

$$\sigma_Z^2 = \frac{17}{448} \quad \text{for each } j = 1 \text{ or } 2$$

Mapping these values back into our original interval I we have

$$E(X_1) = \frac{1}{8}(5a+3b) \quad E(X_2) = \frac{1}{8}(3a+5b)$$

$$\sigma_X^2 = \frac{17}{448}(b-a)^2 \quad \text{for each } j = 1 \text{ or } 2$$

Note that these expressions for the expectation and variance are explicitly independent of m.

As an application, suppose our cost interval I is determined to be I: [\$30, \$50] where a = \$30 and b = \$50. Further, if the "best" expert assessment places the most likely value at appoximately 30% from the lower bound of I, then we could use ρ_1 as our approximating pdf, from which

$$E(X) = \frac{1}{8}(5a + 3b) = \$37.5$$

$$\sigma_X = \$3.9$$

The Triangular Density Function

When little information is available other than subjective estimates on the extreme values of the cost interval I, it is often convenient to apply a triangular density function $\tau(x)$ through the cost range. Classically, $\tau(x)$ has the representation

$$\tau(x) = \alpha^{-1}(1 - \alpha^{-1} \cdot |x|), |x| < \alpha \text{ (Ref. 2)}$$

and is symmetric about the origin in the interval $-\alpha < x < \alpha$. For our purposes, we will define a similar functional form, f_τ, but one that is bounded by $x \ge 0$, and satisfies the boundary conditions, $f_\tau(a) = 0$, and $f_\tau(b) = 0$ with

$$f_\tau(m) = \max_{x \epsilon I} f_\tau(x) \qquad I: [a \le x \le b]$$

Such a function is shown below in Figure 3.

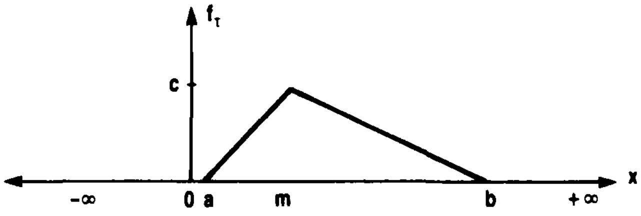


Figure 3. A Triangular Density Function

The probability density function for f_r is defined by

$$f_r = \begin{cases} \dfrac{c}{m - a} (x - a) \text{ if } 0 < a < x < m \\ \\ \dfrac{c}{m - b} (x - b) \text{ if } m < x < b \end{cases}$$

where c, the peak value (modal point) of the triangular density is given by: 2/(b – a). Based on this density, we can compute the first two moments of f_r. They are

$$\mu = E(X) = \int_I x f_r(x)\, dx =$$

$$\begin{cases} \dfrac{1}{3} \cdot (a + m + b) \text{ if m is known} \\ \\ \dfrac{1}{2} \cdot (a + b) \text{ if m is unknown} \end{cases}$$

$$E(X^2) = \int_I x^2 f_r(x)\, dx = \frac{1}{6} \cdot \frac{1}{(b - a)}$$

$$\left\{ \frac{1}{m - a} \cdot (m^3(3m - 4a) + a^4) \right.$$

$$\left. + \frac{1}{b - m} \cdot (m^3(3m - 4b) + b^4) \right\}$$

The cost variance, denoted by σ_X^2 is defined by

$$\sigma_X^2 = E(X^2) - E(X)^2$$

which reduces to

$$\sigma_X^2 = \frac{1}{18} \left\{ (m - a)(m - b) + (b - a)^2 \right\}$$

These simple measures of central tendency are useful for establishing the basis for a cost estimate range when little specific information regarding the nature of a system is available. Measures of expectation, determined by the pdf chosen, inform the analyst where the uncertainty is greatest, skewed to the left or to the right of the modal point. The variance σ^2, can be used to establish a confidence criteria on the bound of a cost estimate range. The next section applies these statistical measures to the problem of establishing a conservative probabilistic cost range based on information obtained from μ and σ^2.

THE CHEBYSHEV BOUND

The integrity of the software cost estimate and any subsequent statistical inference is dependent on the assumption that estimated CPCI size adequately reflects reality. Under this assumption conservative probability statements can be made about the likelihood that the estimated cost range will capture the true cost, that is, to be within some Chebyshev bound. In theory, the Chebyshev bound states that the true value of the cost random variable X differs from the expected cost μ by no more than $k\sigma$ standard deviations, with probability at least equal to $1 - 1/k^2$, $k > 1$:

$$\Pr(|X - \mu| \leq k\sigma) \geq 1 - 1/k^2.$$

No *a priori* assumption about the underlying nature of the cost random variable X is made other than that μ and σ exist. A sketch of this cost range is shown below.

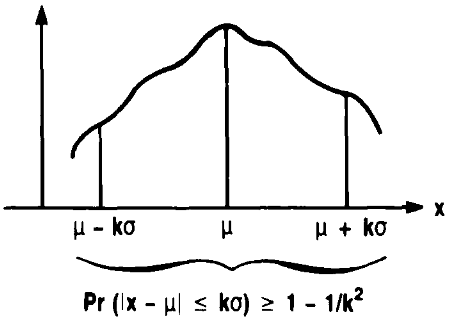


Figure 4. A Chebyshev Cost Range

Applied to a hypothetical system, suppose we determine from a selected density function such as the polynomial or triangular pdf, that

$$E(X) = \$50,000 \text{ and}$$

$$\sigma_X = \$1,500$$

then with an interval length of two standard deviations from the mean there is at least a 75% chance that the true subsystem cost will fall in the cost range $47,000 – $53,000. That is, using the Chebyshev inequality for
k = 2, the

$$\Pr(\$47,000 \leq X \leq \$53,000) \geq 0.75$$

CONCLUSION

The problem of determining a three point approximation to a continuous random variable with an unknown distribution is a popular topic among researchers in the Management Sciences area (Ref. 1). Some mean and variance approximation algorithms are computationally complex and require time consuming computer simulation.

Current research has yet to adequately develop a procedure which models this problem. The informal techniques outlined in this paper support an analytical rationale for assessing uncertainty in a cost estimate. These nonrigorous procedures form the basis of a decision tool that provides the analyst with a method to make conservative probability statements about cost intervals when only subjective technical inputs are available.

REFERENCES

1. Keefer, Donald L. and Bodily, Samuel E., "Three-Point Approximations for Continuous Random Variables", Mgmt. Sci., Vol. 29, No. 5 (1983).

2. Feller, William, "An Introduction to Probability Theory and Its Applications," John Wiley & Sons, N.Y., 1971.

PAUL R. GARVEY is a Member of the Technical Staff at The MITRE Corporation, Bedford, Massachusetts.